数据挖掘与大数据分析
Assignment 2

1. 数据集（20 分）

- 从 UCI dataset repository 中下载以下数据集
  - –（10 分）自行下载一个数据集，要求既包含连续的数值型属性，也包含离散的符号型属性（D_1）；
  - –（10 分）IRIS(D_2), Wine Quality (red vinho verde wine) (D_3), Breast Cancer Wisconsin (Diagnostic) Data Set (D_4)。

  下载以后，仔细阅读数据集的使用说明，理解其用途及每一列数据的含义。

2. 分类器的训练和测试（60 分）

- （15 分）逻辑回归：将 D_4 按照 $|D_{train}| : |D_{test} = 80\% : 20\%$ 的比例进行划分，用 D_{train} 训练一个逻辑回归分类器，用 D_{test} 测试其性能，评价指标可以用 accuracy、precision、recall 和 F_1-measure；
- （30 分）决策树、朴素贝叶斯：分别将 D_1、D_2 按照一定的比例划分为训练集 D_{train} 和测试集 D_{test}（比例自行设定），用 D_{train} 分别训练一个决策树分类器（自选决策树算法）和一个朴素贝叶斯分类器，用 D_{test} 测试其性能，评价指标分别用 accuracy、precision、recall 和 F_1-measure 的 macro、micro 版本；
- （15 分）神经网络、SVM：分别将 D_2、D_3 按一定的比例（自行设定）划分为训练集 D_{train} 和测试集 D_{test}，用 D_{train} 分别训练神经网络（一个输入层、一个隐藏层、一个输出层，隐藏层神经元数目自行设定）和支持向量机模型（SVM 调用 sklearn 包中的实现即可），用 D_{test} 测试训练所得分类器的性能。

3. 撰写技术报告（20 分）

以科技论文的形式撰写 assignment 的技术报告。

- 对于逻辑回归，应给出二分类的 Confusion matrix，并给出各个指标的柱状图；
- 对于决策树、朴素贝叶斯、神经网络、SVM，给出 Confusion matrix，并对其进行分析、解释；
- 对于决策树和朴树贝叶斯，以柱状图或折线图方式对比它们在每一数据集上的性能差异；对于神经网络和 SVM，同样对比其性能优劣；
- 实验部分应对数据集进行介绍，参考文献中给出该数据集的原始出处并在报告正文中第一次出现给数据集的地方添加对文献的引用；
- 对实验结果的呈现，必须以文字形式进行阐述、解释或者说明，不能只是简单地展示结果的图，否则会减分；调整图的大小，使之清晰美观，否则会减分；
- 报告应以正规的书面语言进行客观的阐述，切勿使用口语化的表达方式或使用随意的网络用语；
- 插图应使用矢量图，图、表要添加编号与标题，并在正文中引用其编号；
- 报告中对使用的算法应引用其出处的参考文献，引用格式为用方括号括起来的上标数字形式，按引用的次序依次顺序编号，并在报告末尾添加“参考文献”一节；每一条文献条目中至少应包括作者名，文章标题，期刊名，期号，卷号，出版年月，pp：页码范围，DOI 号或官网的 URL。

4. 必须提交的材料

- 下载的数据集：各个数据集各自存入一个文件中，文件名为程序中使用该数据集时的名称；
- python 的源程序：每个源程序存入一个文件，文件名能体现其作用；
- pdf 版本的技术报告；
- 以上三部分压缩成一个压缩包，以学号 + 姓名对压缩包进行命名。